IL NUOVO CIMENTO Vol. 32 C, N. 2 Marzo-Aprile 2009
DOI 10.1393/ncc/i2009-10378-7

Colloquia: CSFI 2008

Software trigger systems in High Energy Physics

D. Galli

Alma Mater Studiorum, Università di Bologna and INFN, Sezione di Bologna
Bologna, Italy

(ricevuto il 22 Giugno 2009; pubblicato online l'1 Settembre 2009)

Summary. — In order to efficiently collect and store experimental data, the High Energy Physics experiments need a *trigger system* to separate—in *real-time*—interesting experimental events (*signal*) from uninteresting ones (*background*). With the aim of optimising the performance, a modern trigger system splits the data selection task into two or more stages, where the highest-level one is usually implemented as a great deal of software processes (*software trigger*) running *in parallel* on the nodes of a computer farm which can reach the size of 2000–5000 PCs. Data link technologies and operating environment used in the most challenging High Level Triggers of the recent High Energy Physics experiments are here reviewed. In particular, the High Level Trigger of the LHCb experiment at CERN, which has the highest input event rate (1.1 MHz) among the forthcoming experiments, is taken as an example.

PACS 07.05.Hd – Data acquisition: hardware and software.
PACS 29.00 – Experimental methods and instrumentation for elementary-particle and nuclear physics.
PACS 29.85.Ca – Data acquisition and sorting.
PACS 29.85.-c – Computer data analysis.

1. – The trigger systems

In most High Energy Physics experiments the rate at which data are gathered from the detector—*e.g.*, the bunch crossing rate in a colliding beam experiment—is much higher than the rate of the physical events of primary interest. In a typical High Energy Physics (HEP) experiment at the CERN Large Hadron Collider (LHC) the bunch crossing rate is 40 MHz and the event size can be as large as 10 MiB([1]), while the *abundance* of the

([1]) Throughout this paper—following the ISO/IEC standard IEC 80000-13:2008 [1] and the IEEE standard 1541-2002 [2] to avoid ambiguities—the prefixes Ki, Mi, Gi, Ti, Pi, Ei and Zi are used to mean 2^{10}, 2^{20}, 2^{30}, 2^{40}, 2^{50}, 2^{60} and 2^{70}, respectively, preserving for the prefixes k, M, G, T, P, E and Z the original SI meanings of 10^3, 10^6, 10^9, 10^{12}, 10^{15}, 10^{18} and 10^{21}.

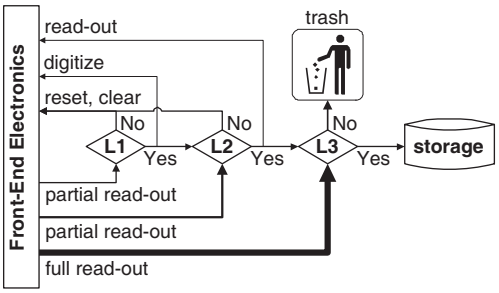

Fig. 1. – A typical layout of a multi-level trigger system.

events of primary physics interest ranges between $1 : 10^6$ and $1 : 10^9$. An experiment could therefore, in principle, acquire several ZiB of data in a year, but the events of interest would contribute only with an amount which ranges from a few TiB to a few PiB.

In order to perform unbiased physical analyses, it would be certainly desirable to save all the acquired data to the mass storage; this option is however practically unfeasible, since the volume of data from the digitisation of all the readout channels is too high to be realistically read out by a data acquisition system (DAQ) and written into a mass storage for later analysis. In practice, the maximum write throughput to the storage system in a HEP experiment has to be considered a *constraint*, fixed by a reasonable budget, and a selection must be operated *on-line* in order to ensure that the maximum number of interesting events is written to the storage media.

The job of a *trigger* equipment is to quickly (real-time) discard uninteresting experimental events while efficiently culling out the most interesting ones in as unbiased a manner as possible. The *trigger* system is so called because an interesting physical event is figuratively supposed to release a lock and set the data acquisition mechanism in action (as it effectively happened in the early trigger equipments).

A trigger system is typically described in terms of *selectivity* (ratio of the trigger rate to the event rate), *efficiency* (fraction of interesting events stored), *rejection* (fraction of background events discarded) and *dead-time* (minimum time interval between two processed events).

2. – Multi-level trigger systems

In the design of a trigger system, four constraints have to be considered: a) the input data rate from the detector, b) the maximum output data throughput to the storage, c) the available computing power for the trigger task and d) the maximum data throughput inside the trigger equipment. Since the real-time processing of the 400 TiB/s of data collected by a modern HEP experiment would exceed both the computing power and the internal data throughput, the modern trigger systems bypass these constraints by splitting the event selection task in two or more stages (fig. 1).

The benefit of the *multi-stage selection* arises from the fact that the most abundant background can be rejected on the basis of a coarse (and fast) selection criterion, involving a very small subset of the data available for the individual event, while the remaining background has a rate sufficiently low to be recognised and rejected by a finer (and more time expensive) selection, involving more data for the single event.

The *coarser* selection is practicable—even if the data rate is very high—because only a *small amount of data* must be processed for the single event and *very simple calculation* (using very simple algorithms) must be performed on each event. The *finer* selection is practicable—even if a large amount of data must be processed for each event and a complex calculation must be performed on the individual event—since the data *rate* has been drastically *reduced* by the *previous filter stage*.

In this line of thought, further third and fourth stages can then be added, which operate a finer and finer selection, by means of more and more time-consuming algorithms which process more and more data per event.

3. – The evolution of the trigger systems in HEP

3˙1. *Historical note*. – In the early HEP experiments, elementary particles left visible tracks which were either directly impressed on a *photographic emulsion* or photographed by a camera equipment: the latter is the case of *cloud chambers* (tracks made of fine water droplets), *bubble chambers* (tracks made of microscopic bubbles in a superheated liquid) and *spark chambers* (tracks made of electric discharges in a gas).

The bubble chamber experiment (1950-70) had not a real trigger equipment. However, in the modern lexicon of trigger systems, we could say that the *low-level* trigger was the device which set in action the camera to record the event (DAQ) during the piston expansion (which made the liquid superheated, thus allowing the bubble development), while the *high-level* trigger were the hired scanners who would look at the film for the bubble tracks representing the events of interest.

The Cronin and Fitch experiment [3] (1964), which allowed the discovery of the *CP* violation, had one of the first real trigger systems. The experiment used, as tracking detectors, spark chambers which needed a fast ($\sim 20\,\text{ns}$) high-voltage pulse to develop sparks, followed by a triggered camera to photograph tracks. The trigger was released by the coincidence of scintillators and water Cherenkov detectors. There was just one trigger level and a long dead-time incurred between two photographs, due to film advance.

As particle physics knowledge developed, frequently occurring events were already well known, while events of interest for ongoing physics research became increasingly rare and had to be selected out of a large number of background events. The advent of electronic devices for data processing—as opposed to older techniques such as bubble chamber photographs—allowed for automatic data processing. Electronic information could therefore be digitised and used for selecting events without human intervention, by means of a full-blown trigger system.

Early trigger equipments had a mere hardware implementation, performed very simple calculation to take the decision and thus their discrimination was very rough. The readout of the event was started by the trigger (DAQ always came after the trigger) and, since it did not foresee the use of pipelines, it caused a large trigger dead-time.

3˙2. *Typical trigger layout*. – A typical modern multi-level trigger system consists in three layers:

– The lowest layer (1st Level or L1) cuts out simple, high-rate background, basing the decision on the feature extracted by the *partial* readout of a *single* sub-detector. It is implemented in the hardware, by means of custom electronics.

  In order to perform fast calculation, very often the algorithmic computations are replaced by *look-up tables* (LUTs), which yield the results in a smaller number of

clock cycles. When the computing time exceeds the interval between two subsequent events, in order to avoid *dead-times*, memory *pipelines* (*i.e.* circular buffers) are used to keep the physical event stored during the time needed by the trigger to come to a decision.

In recent L1 trigger implementations powerful ASICs([2]) and, more and more often, programmable FPGAs([3]) are used.

In a few cases even hardware implementations of *neural networks* are employed in the first-level trigger, like the Adaptive Solutions CNAPS chip at the H1 experiment at DESY, and the Intel ETANN chip at the CDF experiment at Fermilab.

– The intermediate layer (2nd Level or L2) is able to make a more refined analysis of the event by matching the feature extracted by several sub-detectors. It often has a hybrid hardware/software implementation and sometimes makes use of DSPs([4]), like the Inmos Transputer in the Zeus experiment and the Analog Device Sharc in the Hera-B experiment [4].

  Sometimes (this is the case, *e.g.*, of the Hera-B experiment [4] at DESY and of the Atlas experiment [5] at CERN) the 2nd level trigger processes the so-called *regions of interest* (ROI). Following a L1 accept signal, the detailed L1 information is forwarded to a ROI controller, which uses the 1st level trigger information details to locate portions of the detector it is interested in (*e.g.*, a portion of a calorimeter hit by a particle jet) for a spatially limited full readout. The L2 trigger performs a full event reconstruction limited to the ROI. If the event passes the L2's requirements, the results of L2, along with the full detector bandwidth, are transmitted to the following trigger level. If, on the contrary, the event is found to be lacking, the rest of its data need never to be transmitted, thus reducing the overall bandwidth requirements of the network.

– The highest layer (3rd Level, L3 or HLT, High Level Trigger) performs a full event reconstruction of the events which have passed the previous trigger layers. It is usually implemented through a number of software processes (up to 15000–40000) which execute in parallel on the nodes of a computer farm which can reach the size of 2000–5000 PCs (typically one process per CPU core).

3˙3. *Recent trends in trigger systems.* – Experimental background is more efficiently rejected and signal kept if the full granularity of the detector can be used when making the trigger decision. In the extreme this implies to have only one trigger stage.

The advances in technology help moving toward this target from the two opposite sides.

On the one hand, they are pushing the use of the HLT downward. The increasing speed and the decreasing cost of COTS (commercial off-the-shelf) computers and network

([2]) Application-Specific Integrated Circuit, an integrated circuit customised for a particular use, rather than intended for general-purpose use.
([3]) Field-Programmable Gate Arrays, standardised semiconductor devices that can be configured by the customer after manufacturing. They can be programmed by writing a software code in a relatively high-level programming language (such as VHDL), which is then compiled by a computer to produce the binary file that has to be loaded into the chip.
([4]) Digital Signal Processor, a specialised microprocessor designed specifically for digital signal processing, generally in real-time computing.

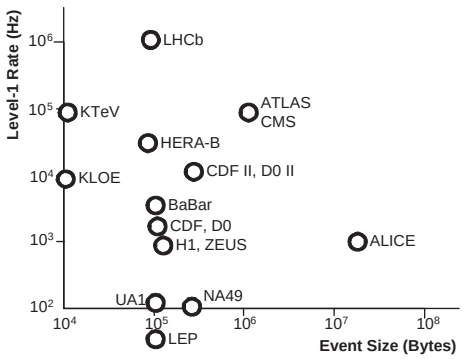

Fig. 2. – Low-level trigger rates and event sizes in several HEP experiments.

devices allow the HLT to work at higher and higher rates and thus allow (and make convenient) to move a larger bandwidth of the detector data into the HLT (fig. 2). In other words, they allow to migrate the HLT decision to a lower level.

On the other hand, the advances in technology are pushing the use of the L1 upward, by allowing to use the L1 trigger at a higher level. The improvement in programmable FPGAs allows to software-code even complex calculation in silicon, achieving event rejection rates never before associated with such good physics efficiency, with the possibility of modifying and improving the trigger logics without rebuilding the electronic circuits (the line between hardware and software triggers is now blurring more and more).

As a result, the traditional Level 2 is disappearing (except for software ROI processing). In a few cases, in the design of new experiments, the software trigger is foreseen as the only trigger stage (*e.g.*, in the ILC project [6] and in a LHCb upgrade project [7]).

4. – Hardware trigger versus software trigger

Software triggers have several advantages with respect to hardware ones: among them *flexibility* (the selection rules for the events can be changed simply by modifying a software code), *scalability* (the processed event rate can be increased simply by increasing the number of the PCs and the network switch size), *cost* (commodity components used in software triggers are very cheap and their prices rapidly drop), *maintainability* (widespread commodity interfaces will continue to be available on the market), *upgradability* (the software triggers can profit from the rapid development undergone by commodity components).

The drawback of the software triggers is the *variable latency*: a software trigger can take different amounts of time to reach the decision for different events. This makes it difficult to use a software trigger for other but the last trigger level, since data must be kept in a buffer during trigger processing and the buffers can be filled completely if a trigger decision requires too much time.

5. – Trigger and DAQ

In the early HEP experiments the DAQ always came after the trigger: data were read out from the electronics, assembled and formatted to be put in the mass storage only after the trigger decision was taken.

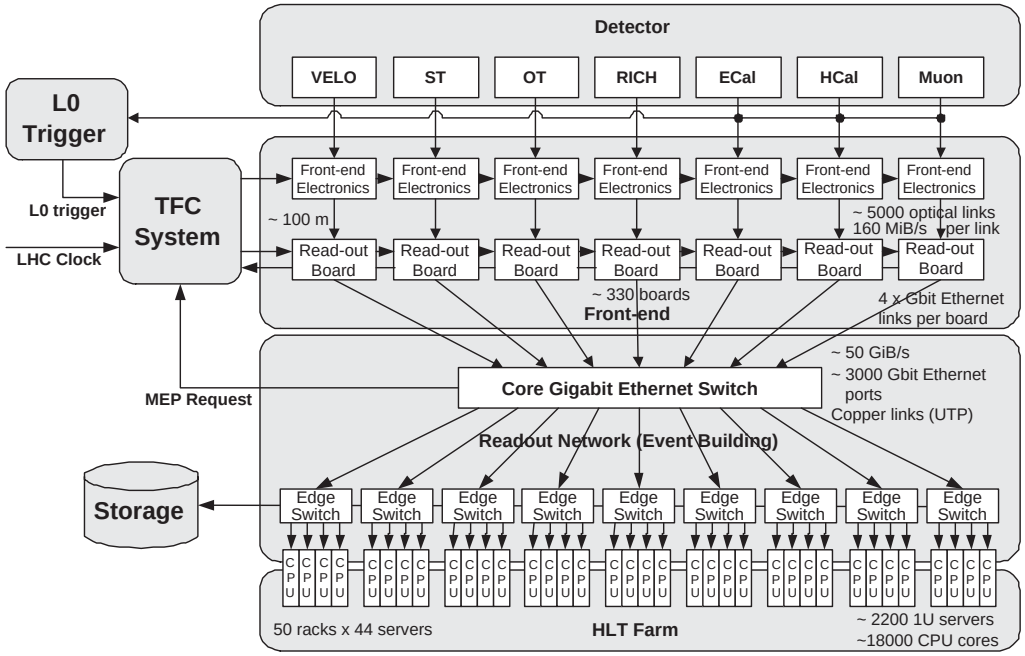

Fig. 3. – Layout of DAQ and HLT of the LHCb experiment at CERN.

In the effort to delay the trigger decision further and further, the DAQ now frequently comes in the middle of the trigger—after a hardware-based L1 but before a software-based HLT. Along the way the event must be built: the data packets coming from the different sub-detector elements and pertaining to the same physical event have to be assembled together in a complete event (*event building*).

The event building task requires, of course, not only data processing technology but also data link technology. The observed trend in DAQ/trigger systems is to move the event building task from hardware to software and from proprietary to commodity link technology.

In the LHCb experiment [8, 9] (fig. 3), for example, after the L1 trigger decision, data fragments are read-out from the sub-detector elements, through $\sim$ 5000 optical links, by $\sim$ 330 *read-out boards* (electronic boards, named TELL-1) which then chunk the event fragments together in groups of 15 (MEP, Multi Event Packet) pertaining to 15 subsequent events which have passed the L1 filter (ME, Multi Event). MEPs are then pushed directly into the *read-out network*, which is a high-end commodity Gigabit Ethernet network ($\sim$ 3000 Gigabit Ethernet ports), which support an average aggregated network throughput of $\sim$ 50 GiB/s. The need to chunk events in MEs arises from the optimisation of the Gigabit Ethernet performance. For each ME, a *Trigger and Fast Control* system (TFC) broadcasts to all the readout boards the destination IP address of the farm PC designate to process that ME so that all the readout boards can send the MEPs pertaining to that ME to the designate node, which then performs the event building of all the events in the ME and takes, for each event, the trigger decision. TFC is responsible for the balancing of the computing load among all the farm nodes.

6. – Software trigger technology

The rate of the physical events reaching the HLT is usually larger by 3-4 orders of magnitude with respect to the rate of the HLT event processing with the aim of taking a trigger decision, thus computing must be performed in *parallel*.

Since the physical events are wholly independent of each other, they can be simply distributed among a number of unconnected processes, running on the computer farm, without the need of interprocess communication (*loosely coupled parallelism*): each process, separately, performs the event reconstruction and takes the trigger decision related to a specific event, without the need of communication with the other processes.

In the LHCb experiment MEs are distributed among $\sim$ 2200 PCs, for a total of $\sim$ 18000 CPU cores. Each PC runs eight HLT processes (as many as the number of CPU cores in the PC), one event builder process (which assemblies together the received MEPs to build a ME and distributes the events in the ME among the trigger processes) and one writer process (which forwards the accepted events to the storage). Inside a given farm PC, the involved processes exchange data with each other by swapping *descriptors*, which point to the real data, while the real data themselves are kept in a shared memory area, and thus not really moved or copied from a process to another one in the same PC.

6˙1. *Rightsizing a trigger farm.* – Since a software trigger is not designed to have a fixed latency, we can think in terms of average values in sizing a trigger farm (*i.e.* in evaluating the minimum required number of PCs).

The PC farm has to be sized in order for the average time spent for the selection algorithm, $\langle T_s \rangle$, to be less than the average period which separates the incoming of two subsequent events into the same trigger node, N_{cpu}/ν_{input}. So, in order for the farm to be able to process the required rate, it must be: $N_{cpu} \geq \langle T_s \rangle \nu_{input}$. As an example, if the mean time required by the selection algorithm is $\langle T_s \rangle = 2$ ms and the input rate is $\nu_{input} = 1.1$ MHz, then the number of CPUs must be $N_{cpu} \geq 2200$.

6˙2. *Farm operating systems.* – In the last years the hardware platform used to implement a HLT farm has rapidly moved from RISC servers to Intel-compatible PCs. If in the BaBar experiment [10] at SLAC (designed in 1995) still Sun SPARC servers are used, the Intel-compatible PCs are already used in both CDF [11] and DØ [12] experiments at Fermilab and all the four LHC experiments at CERN employ Intel `x86_64` PCs.

A number of different operating systems can run on Intel-compatible PCs, among them: Linux, Solaris, Lynx-OS, FreeBSD, Mac-OS, MS Windows, etc. The Linux operating system seems now to stand out. Few exceptions in the recent years are the DØ experiment [12], which ran MS Windows on its L3 Farm (but switched to Linux in Run II) and the CDF experiment [11], which runs VxWorks for a specific task (data transfer from VME readout boards to ATM network).

6˙3. *High-speed data link technologies.* – Similarly to other technologies, also data link technologies adopted in HLTs are moving from custom technologies to widespread commodity and open standard technologies, as soon as they become available.

The Hera-B experiment [4] at DESY, designed in 1995, used the Sharc links (proprietary, by Analog Devices, the same manufacturer of the employed DSPs) in the 2nd-level trigger, but in the HLT used Fast Ethernet links. The DØ experiment [12] uses Fast Ethernet and Gigabit Ethernet. The CDF experiment [11], designed a few years before DØ, uses ATM (a promising open standard technology at the time of the DAQ design).

At the LHC, the CMS experiment [13] uses the Myrinet links (proprietary, by Myricom) for the FED Builder and Gigabit Ethernet for the other links. The other experiments (Atlas, LHCb and Alice) employ Gigabit Ethernet links.

Next-generation experiments will probably choose among the Ethernet technology—already available, in the 10 Gbit/s flavour, on optical fibre and now also on copper UTP cables and probably available, in the year 2010, in the 100 Gbit/s flavour—and a new competitor technology, InfiniBand, already available in 10 to 40 Gbit/s flavour which could have several advantages over Ethernet (low-latency, remote-DMA, etc.).

7. – Conclusion

The trigger systems are used in High Energy Physics experiments to separate in real-time interesting experimental events from background. More and more often trigger filter stages have a software implementation, consisting in a number of processes (up to 15000–40000) running in parallel on a computer farm (loosely coupled parallelism), which reconstruct the physical events from the detector signals and take the final acceptance/rejection decision. High-performance commodity computing and link technologies are very attractive as a hardware platform for such systems.

REFERENCES

[1] IEC and ISO, *Quantities and units - Part 13: Information science and technology*, URL: http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=31898.

[2] Frysinger J. R. *et al.*, *IEEE Trial-Use Standard for Prefixes for Binary Multiples*, ISBN: 0-7381-3385-8, URL: http://ieeexplore.ieee.org/iel5/8450/26611/01186538.pdf.

[3] Christenson J. H. *et al.*, *Evidence for the 2 π Decay of the k_2^0 Meson*, *Phys. Rev. Lett.*, **13** (1964) 138, URL: http://prola.aps.org/abstract/PRL/v13/i4/p138_1.

[4] Egorytchev V. *et al.*, *Architecture of the HERA-B Data Acquisition System*, *IEEE Trans. Nucl. Sci.*, **50** (2003) 859, URL: http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=01221888.

[5] Kordas K. *et al.*, *The ATLAS Data Acquisition and Trigger: concept, design and status*, in *Proceedings of the 10th Topical Seminar on Innovative Particle and Radiation Detectors*, *Nucl. Phys. B - Proc. Suppl.*, **172** (2007) 178, URL: http://www.sciencedirect.com/science/article/B6TVD-4PYS2WC-1V/2/553caf429e610ecdb313d240d1431f14.

[6] Eckerlin G. and Le Du P., *Trigger/Data Acquisition Issues and Challenges for the Next Generation of Experiments at the Future International Linear Collider*, in *Proceedings of IEEE Real-Time Conference, 2007 15th IEEE-NPSS*, ISBN: 978-1-4244-0867-2, URL: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=4382831.

[7] Parkes C., *The LHCb Upgrade*, in *Proceedings of the 2007 Europhysics Conference on High Energy Physics*, *J. Phys.: Conf. Ser.*, **110** (2008) 092021, URL: http://www.iop.org/EJ/article/1742-6596/110/9/092021/jpconf8_110_092021.pdf.

[8] Alessio F. *et al.*, *LHCb Online Event Processing and Filtering*, in *Proceedings of 2007 CHEP Conference, Victoria, B. C., Canada*, *J. Phys.: Conf. Ser.*, **119** (2008) 022003, URL: http://www.iop.org/EJ/abstract/1742-6596/119/2/022003/.

[9] Augusto Alves A. jr. *et al.*, *The LHCb Detector at the LHC*, *Journal of Instrumentation*, **3** (2008) S08005, URL: http://iopscience.iop.org/1748-0221/3/08/S08005.

[10] Jacobsen R. *et al.*, *The BABAR Event Building and Level-3 Trigger Farm Upgrade*, in *Proceedings of 2003 CHEP Conference, La Jolla, CA, USA, CHEP-2003-MOGT003*, SLAC-PUB-9666, *Proc. eConf C* 0303241 MOGT003 (2003), arXiv:physics/0305136, URL: http://arxiv.org/abs/physics/0305136.

[11] ANIKEEV K. *et al.*, *Event-Building and PC Farm based Level-3 Trigger at the CDF Experiment*, *IEEE Trans. Nucl. Sci.*, **47** (2000) 65, URL: http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=846119&isnumber=18359.

[12] ANGSTADT B. *et al.*, *Ethernet-based data acquisition system for the DØ experiment at Fermilab*, in *Proceedings of 2003 IEEE Nuclear Science Symposium, Portland, Oregon, USA*, ISBN 0-7803-8257-9, URL: http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=1351854&isnumber=29713.

[13] BAUER G. *et al.*, *CMS DAQ Event Builder Based on Gigabit Ethernet*, *IEEE Trans. Nucl. Sci.*, **55** (2008) 198, URL: http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=04448531.